学号姓名_课程研学报告（一）

学号： **姓名：** **日期：**

课程研学报告 撰写格式要求： 正文宋体 5 号，A4，1.25 行距

课程报告撰写内容：

一、建立文档库

1. 构建方法：将“关于 XXXX 搜索引擎的学习报告”中第三部分选择的学术性论文的参考文献列表内的每篇参考文献的题目，作为一个文档(doucment)，所有参考文献的题目，构成文档库。

例如，论文“**Research on Development Architecture of WeChat Platform**”的文档库如下所示：

D1: Building a framework of rapid development of WeChat public platform

D2: The theory adout the software software architecture

… …

REFERENCES

[1] Guanghao Liang. Building a framework of rapid development of WeChat public platform [J]. Volkswagen technology, 2015, (10) : 134-137 + 155.

[2] Jiancheng Zhao,Xueping Zhou. The theory adout the software software architecture [J].Computer programming skills and maintenance, 2008, (16) : 1-2.

… …

2. 文档数量：文档库的文档数量必须大于 30。

3. 补充文档的方法：如果论文的参考文献数量少于 30，按发表年限降序顺序，选择最新发表的一篇参考文献，重新获取该参考文献的原文，将该原文中的参考文献，补充到文档库中。

二、建立索引

1. 利用课程教材中第 1-2 章中介绍的方法，对文档库建立倒排索引。

2. 在哈希表、二叉查找树、B 树和 B+树中，选择一种数据结构，提供对倒排索引结构中的字典部分的信息项(term)的快速查询，并写出相应的查询算法。

三、建立文档向量

将倒排索引结构中的字典作为“词袋”，对文档库中的所有文档，建立文档向量。

四、建立查询和查询向量

将“关于 XXXX 搜索引擎的学习报告”中第三部分选择的学术性论文的题目，作为查询 q，如 query ： **Research on Development Architecture of WeChat Platform**，并建立查询向量。

2017 3rd IEEE International Conference on Computer and Communications

Research on Development Architecture of WeChat Platform

五、利用向量模型，计算查询 q 与所有文档的相关度。

1. 计算所有文档向量和查询向量的 tf, wf, de, idf 的值。计算方法，参考第 6 章中描述的方法，如下图所示。

	query					document			
word	tf	wf	df	idf	q_i = wf-idf	tf	wf	d_i = normalized wf	$q_i \cdot d_i$
digital			10,000						
video			100,000						
cameras			50,000						

Table 6.1

【说明】：有关 wf 的计算，每个同学可以在第 6 章介绍的所有计算方法中，自行任选一种计算公式，且需要在报告中明确给出该计算公式。

2. 计算查询 q 与所有文档的相关度，按相关度从高到低排序后输出。

六、利用基于概率模型的语言模型，计算查询 q 与所有文档的相关度。

1. 计算并建立所有文档向量和查询向量的语言模型。参考第 12 章中描述的方法。如下图所示。

Model M_1		Model M_2	
the	0.2	the	0.15
a	0.1	a	0.12
frog	0.01	frog	0.0002
toad	0.01	toad	0.0001

【说明】 ：有关先验概率 P 的计算公式，每个同学可以在第 11 章 11.3.3 节介绍的计算方法中，自行任选一种计算公式，且需要在报告中明确给出该计算公式。

11.3.3 Probability estimates in practice

2. 计算查询 q 与所有文档的相关度，按相关度从高到低排序后输出。计算公式：

$$P(d|q) \propto P(d) \prod_{t \in q} ((1-\lambda) P(t|M_c) + \lambda P(t|M_d)) \tag{12.12}$$

七、对比和分析两种相关度计算的结果，谈谈个人对信息检索相关度计算方面的看法。